**Module Assignment**

**Module 6**

**QMB-6304 Foundations of Business Statistics**

**#Debjani Sarma**

**rm(list = ls())**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load the file “6304 Module 6 Assignment Data.xlsx” into R. This file contains information on 1338 instances of an adult being hospitalized somewhere in the United States. The variables included the patient’s age, body mass index (bmi), whether or not they were a smoker, and the total final charges for hospital care submitted to the patient or a third-party payer. This is your master data set.

**> hospital.data=rio::import("6304 Module 6 Assignment Data.xlsx")**

1. Using the numerical portion of your U number as a random number seed, take a random sample of 150 cases from the full data set using the method presented in class. Convert smoking status to a factor variable. This will be your primary data set for analysis.

**> set.seed(24173877)**

**> my.data=hospital.data[sample(1:nrow(hospital.data),150),]**

**> my.data$smoker = as.factor(my.data$smoker)**

**Analysis**

Using your primary data set:

1. Show the results of the str() command.

**> str(my.data)**

**'data.frame': 150 obs. of 4 variables:**

**$ age : num 41 30 20 42 33 30 63 61 33 37 ...**

**$ bmi : num 40.3 24.1 33 30 33.4 ...**

**$ smoker : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 2 2 2 ...**

**$ charges: num 5709 4032 1880 22144 6654 ...**

1. Conduct a full regression analysis including all variables and using the “charges” variable as the dependent variable.

**> attach(my.data)**

**> my.regmodel = lm(charges ~ age + bmi + smoker, data = my.data)**

1. Show your model output. Interpret the beta coefficients in your output in terms of the variable’s estimated impact on the y. Include an appropriate discussion of the beta coefficient p values.

**> summary(my.regmodel)**

**Call:**

**lm(formula = charges ~ age + bmi + smoker, data = my.data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-10520 -3459 -1971 1076 22047**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -9276.56 3136.15 -2.958 0.00361 \*\***

**age 310.25 39.64 7.826 9.36e-13 \*\*\***

**bmi 198.73 94.14 2.111 0.03648 \***

**smokeryes 22409.87 1385.88 16.170 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 6610 on 146 degrees of freedom**

**Multiple R-squared: 0.6952, Adjusted R-squared: 0.6889**

**F-statistic: 111 on 3 and 146 DF, p-value: < 2.2e-16**

**Beta-coefficient**:

**Intercept**: It represents that the final charges is -$9276.56, when all the independent variables are 0. In practical sense, having a -$9276.56 medical charge is not meaningful as having 0 for age, BMI is not realistic.

**p-value**: With a p-value of 0.00361, the intercept is statistically significant. It means we are rejecting the null hypothesis (intercept = 0) and accepting the alternate hypothesis (intercept !=0).

**Age**: It represents that if age is increased by 1 year, the final medical charge increases by $310.25.

**p-value**: With a p-value 9.36e-13, age is highly significant, suggesting that age has strong relationship with final medical charges. This indicates that as age increases, the medical charges are likely to increase as well.

**BMI**: It represents that if BMI is increased by 1 unit, the final medical charge increases by $198.73.

**p-value**: With a p-value 0.03648, BMI is significant, suggesting that BMI has a relationship with final charges. This indicates that as BMI increases, the medical charges are likely to increase as well.

**smokeryes**: It represents that being a smoker is associated with an increased in the final medical charges by approximately $22409.87 compared to non-smokers.

**p-value**: With a p-value < 2e-16, this coefficient is highly significant, suggesting a strong relationship with final medical charges.

1. Report the confidence interval for each beta coefficient in your model.

**> confint(my.regmodel)**

**2.5 % 97.5 %**

**(Intercept) -15474.68273 -3078.4340**

**age 231.90215 388.6009**

**bmi 12.67578 384.7864**

**smokeryes 19670.88528 25148.8462**

1. Determine and state whether your model appears to be in conformity with the LINE assumptions of regression. Show appropriate graphics where needed to justify your conclusions.

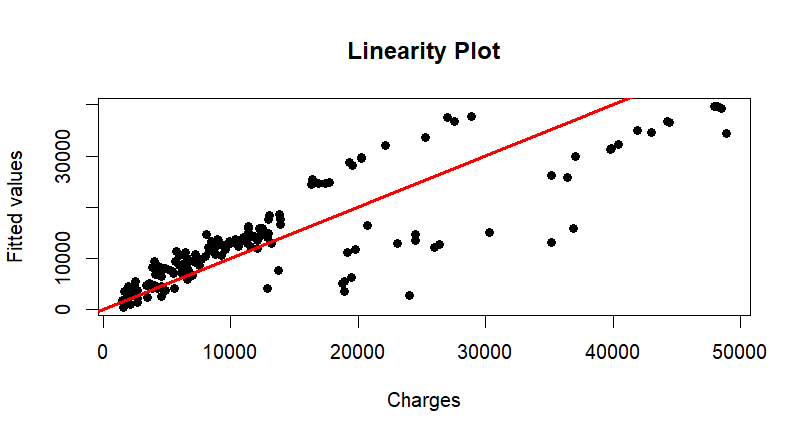
**> #Assumptions of LINE**

**> #Linearity:**

**> plot(charges, my.regmodel$fitted.values, pch=19, main = "Linearity Plot",**

**+ xlab="Charges", ylab="Fitted values")**

**> abline(0,1,lwd=3, col="red")**

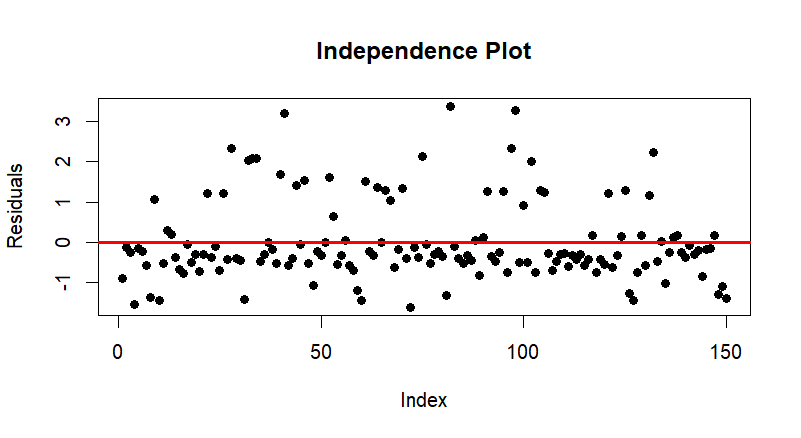


From the linearity plot, we can say that the relationship between the charges and fitted values appears to be nearly linear. However, there are some deviations at the higher ends, which suggests the relationship might not be perfectly linear at the extreme values.

**> #Independence:**

**> plot(scale(my.regmodel$residuals),pch=19, main = "Independence Plot", ylab="Residuals")**

**> abline(0,0, lwd=3, col="red")**



From the independence plot, there seems to be no obvious patterns in the residuals, indicating that we are in conformity with the assumptions of independence.

**#Normality:**

**> qqnorm(my.regmodel$residuals, pch=19, main ="Normality plot")**

**> qqline(my.regmodel$residuals, lwd=3, col="red")**

**> hist(my.regmodel$residuals, main="Histogram of residuals", col="pink", probability = TRUE,**

**+ xlab = "Residuals")**

**> curve(dnorm(x,mean(my.regmodel$residuals),sd(my.regmodel$residuals)),**

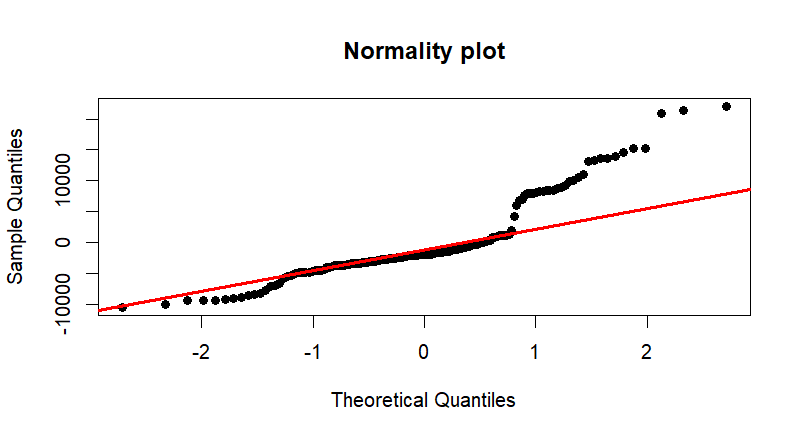
**+ from=-20000, to = max(my.regmodel$residuals), lwd=3, col="blue",add=TRUE)**

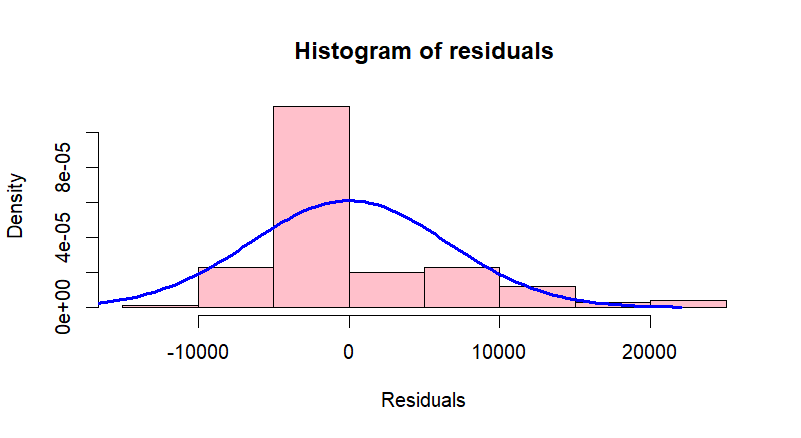
**> moments::skewness(my.regmodel$residuals)**

**[1] 1.2596**

**> moments::kurtosis(my.regmodel$residuals)**

**[1] 4.343604**





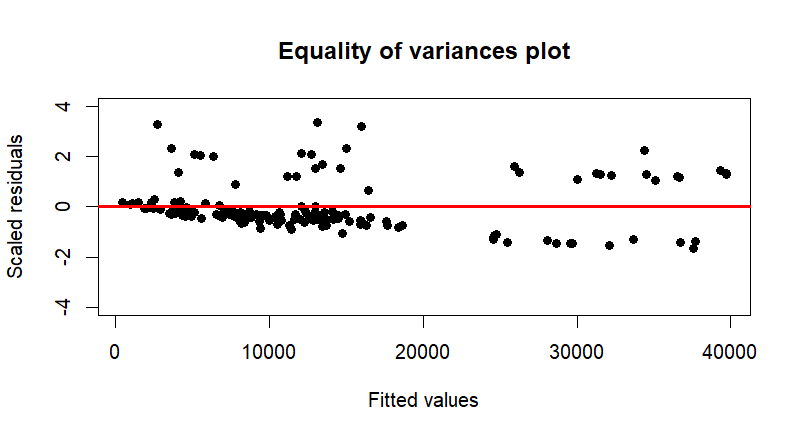
The normality plot shows deviation from normality, especially at the higher end values, suggesting the residuals are not perfectly normally distributed. The skewness of 1.2596 and kurtosis of 4.343604 indicates positive skewness and leptokurtosis. The histogram of residuals verifies these findings, indicating that the residuals are not normally distributed. These values suggests that the assumption of normality is violated.

**> #Equality of variances:**

**> plot(my.regmodel$fitted.values, rstandard(my.regmodel), pch=19, main="Equality of variances plot",**

**+ xlab="Fitted values", ylab="Scaled residuals", ylim=c(-4,4))**

**> abline(0,0,lwd=3, col="red")**



The equality of variances plot shows that there may be some heteroscedasticity, as residuals are more spread out at higher end of the values. This violates the assumptions of equality of variances.

The model conforms with the assumptions of linearity but might not be perfectly linear at the extreme values. The model also conforms with the assumptions of independence. However,it slightly violates the assumptions of normality and equality of variances.

1. Determine whether any of the data points in your reduced data set have a high leverage in influencing the plot of the regression. Show appropriate analytics to support your conclusion. Also, report the observations from your reduced data set (if any) which have such high leverage.

**> #Determining the leverage points**

**> leverages = hat(model.matrix(my.regmodel))**

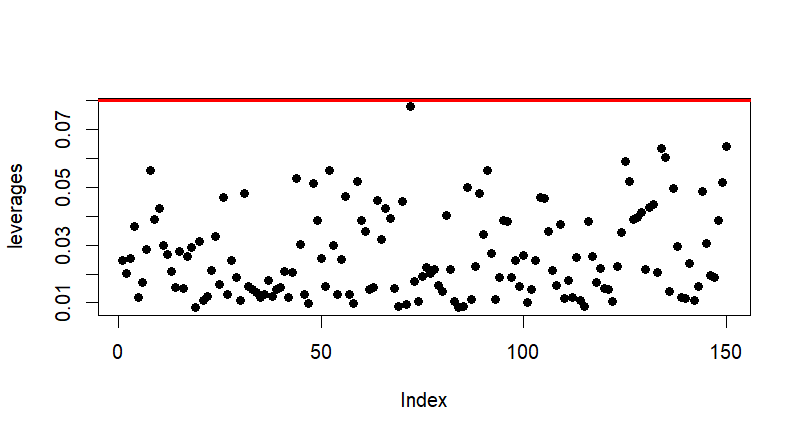
**> plot(leverages, pch = 19)**

**> abline(3 \* mean(leverages), 0, col="red", lwd=3)**

**> my.data[leverages>(3\*mean(leverages)),]**

**[1] age bmi smoker charges**

**<0 rows> (or 0-length row.names)**



**Interpretation** : The leverage plot shows the leverages of all data points with threshold of 3 \* (mean leverage). The result of which shows that no data points exceed the threshold, including the mean of the observations having a high leverage.

**Conclusion** : There are no high leverage data points in the reduced dataset that significantly influences the regression model.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is “Assignment 5” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**